ASSESSING AND IMPROVING DATA QUALITY IN OPEN SPATIAL DATA: A CASE STUDY WITH ANAC DATA

Vincenzo Nardelli[1], Niccolò Salvini[2]

[1] Department of Economics, quantitative methods and business strategies, University Milano Bicocca,
(e-mail: v.nardelli2@campus.unimib.it)

[2] Department of Healthcare Management, (e-mail: niccolo.salvini@unicatt.it)

ABSTRACT: In this paper, we focus on assessing data quality in the context of open anti-corruption data, using data from the National Italian Anti-Corruption Authority (ANAC). The open data movement promotes governments to publish data sets to enhance transparency and accountability, which has been particularly beneficial in combating corruption. Nonetheless, open data is not exempt from challenges, one of which is data quality. We investigate missingness of the data to determine if it is missing at random, not at random, or completely at random. We then present a data quality algorithm, specifically designed for ANAC data, to sanitize errors. To further investigate the phenomena of missingness, we enriched the dataset with socio economic indicators at municipality level coming from official sources (such as ISTAT and Ministry of Economy and Finance, etc.). Finally, we use modelling to determine the factors that contributed to the missingness . An important addition to the classical modeling approaches widely used in literature is to assess if the missingness depends also on the geo-localization of the municipality. This is carried on testing whether it exists an autocorrelation on residuals not explained by classical methods. Our results indicate that addressing missing data with the proposed methodology can lead to more accurate and reliable data for anti-corruption assessments. This study contributes to literature on data quality assessment and provides insights into the challenges and potential solutions to missing data in open data initiatives exploiting spatial statistics techniques.

KEYWORDS: SWORD, data quality, spatial modelling.